# CHAPTER VII - TN 36: THE MEANING OF R2 IN USING ANALYSIS OF VARIANCE TO EXPLAIN PARTICIPATION IN RECREATION ACTIVITIES

By M.F. Goodchild

ABSTRACT

The probability of an individual's participation in a recreation activity has been modelled as a linear combination of socio-economic variables, or traits. Such models are calibrated by using as the dependent variable a binary observation of whether each individual did or did not participate in the activity.

The $R^2$ statistic for these models is shown to be dependent only on the statistical relationship between probability and binary observation. $R^2$ in no way indicates the validity of the model. Estimates of coefficients are unbiased but, when OLS is used, standard errors in coefficients are overestimated (note that using GLS resulting in lower standard error is a topic of the Smith and Cicchetti appendix to this CORDS volume).

An alternative index of structural error influence is proposed and its use is demonstrated by simulation.

## BACKGROUND

CORDS TN 6, 12, 13 20, and 29 examined the use of regression with dummy variables (also called analysis of variance in CORDS) as a method of modelling participation in recreation activities. Briefly, it was argued that in order to project demand for a given activity in a community, a model must incorporate shifting socioeconomic patterns. Rising incomes and aspirations create changes in leisure habits which in turn are reflected in varying patterns of demand for both indoor and outdoor activities.

The model proposed characterizes an individual by a bundle of socioeconomic variables, and postulates that some combination of these determines the probability of participation in a given activity by that individual:

$p = a_0 + \Sigma_i a_i x_i + e$ i=1 to m

WHERE ε is an error term, to be discussed below, $a_0$ and the $a_i$ is a vector of constants, $x_i$ is the value of the ith socioeconomic variable, and m is the number of such variables.

In many cases the values of $x_i$ will be restricted to 1 or 0, corresponding to the individual's possession of the $i^{th}$ socioeconomic trait, or his lack of it, respectively. TN 12 uses the terminology of analysis of variance, since all socioeconomic variables are considered as binary traits by allocating a dummy variable to indicate presence or absence of a level of an attribute (e.g., being in a given age range is a variable). The terms and symbols used in this paper are those of multiple regression. This is done since many of the conclusions apply equally to independent variables measured on continuous scales. The Time Series Processor (TSP) was used for all simulations, in conjunction with small FORTRAN routines being used for data generation and post-TSP analysis.

To predict and project participation levels in a community as a whole, the model is simply summed over the members of the community;

$E(n) = \Sigma p_{i,j}$

$= \Sigma_j a_0 + \Sigma_j \Sigma_i a_i x_{i,j} + \Sigma e_j = Na_0 + \Sigma a_i n_i + \acute{\varepsilon}$ j=1 to N and i=1 to m

WHERE E(n) is the expected number of participants,

N is the population of the community,

$n_i$ is the population having the ith trait (e.g., a level of some variable), and

j specifies a single individual.

TN 12 illustrates the calibration of the model against the results of a survey with data on

individual participation, and the use of the model in projecting recreation demand. The present paper examines a series of problems in the use of the model, and in the interpretation of the results, concentrating on the disappointingly low $R^2$ values, or measures of the model's fit to reality, that are customarily obtained.

## THE BASIS OF THE PROBLEM

In calibrating the model, the dependent variable is not an observation of the probability p that an individual will participate, but an observation of the individual's actual behaviour. If the individual was observed to participate, the dependent variable is given the value of 1, and if not, 0. The relationship between this dichotomous variable and the probability p, is statistical: the probability that a 1 will be observed in a single trial is p, and that a 0 will be observed, l-p.

The effect of calibrating with the dichotomous variable, which will be referred to as y, rather than p, can be thought of as the addition of another error term to the model. The conventional error in p, which was introduced in Equation 1 as *e*, will be referred to as structural error, while the second source introduced by using y as the dependent variable will be called statistical error. While structural error is an empirical quantity, statistical error should be predictable (*e* reflects how much the model is in error for predicting an individuals true probability—it is a correction to an imperfect fit of a general model to individuals).

Suppose that structural error is absent, so that the entire error in the model is due to the statistical component. Consider a very large sample, and let the distribution of p values be such that the probability of finding a case in the interval p to p + dp is f(p)dp. Since p must lie in the range 0 to 1 it must be true that:

$\int f(p)dp = 1$

WHERE the integration is from 0 to 1.

The probability of observing y to be 1 is p, and 0, 1-p. With a sample of N, the expected number of observations for which y = 1 will be Npf(p)dp, and for y = 0; N(1-p)f(p)dp. So the observed sum of squared deviations between y's and p's, or between observed and predicted values of the dependent variable, will be: Nf(p) p(1-p)dp

The total sum of squares for all such intervals will be: $\int Nf(p)\ p(1-p)dp$ (integrating from 0 to 1) which is thus the sum of squares of y's about p's.

The total sum of squares for the dependent variable y about its own mean is:

$\int Npf(p)dp - (\int Npf(p)dp)^2/N$ (integrating from 0 to 1)

Thus the value of $R^2$ that will be observed in the absence of structural error has an expected value of:

$(k2 - k1^2)/(k1 - k1^2)$ WHERE $k1 = \int pf(p)dp$ and $k2 = \int p^2 f(p)dp$

(integrating from 0 to 1)

So the asymptotic value of $R^2$ due to the statistical error component can be predicted from the first and second moments of the distribution of p. Consider a simple example. Suppose that the model predicts only two values of p. 0 and 1. That is, under certain socioeconomic conditions, individuals will certainly participate, and under all other conditions will certainly not.

Suppose that the probability of either condition occurring is 1/2.

Then $k1 = 1 * 1/2 + 0 * 1/2 = 1/2$

$k2 = 1^2 \times 1/2 + 0^2 \times 1/2 = 1/2$

$R^2 = 1$

Since p is restricted to 0 and 1, the y's must be respectively 0 and 1, there is no statistical error, and the fit is perfect.

Now let the two equally likely combinations of independent attributes give rise to p values of 1/3 and 2/3. This time statistical error is present, and $R^2$ deteriorates dramatically;
k1 = 1/2; k2 = 5/18; $R^2$ = 1/9

The foregoing discussion is of course limited to the asymptotic case. In principle, it is possible to calculate the distribution of $R^2$ for samples of limited size, but the tractability of the problem will depend very much on the distribution of p, and therefore on the empirical $x_i$ and $a_i$. For this reason the developments which follow are based on simulation of a few realistic cases, rather than on general mathematical analysis.

NATURE OF THE STRUCTURAL ERROR TERM

Many factors influence the likelihood of an individual's participation in an activity besides the socio economic traits $x_i$ included in the model. These include the effects of varying levels of the supply of opportunities for recreation (see TN 29), of the individual's own learning process, and of attitudes and perceptions. Furthermore, the model may not include interaction effects between two or more traits. Thus, it may be assumed that education is uniformly influential as a trait regardless of the presence or absence of other traits. But education may take part in interactive effects with other factors. If, for example, university education affects a certain participation probability only when coupled with high income, its influence will go undetected and appear in the error term, along with any other such interaction effects, or may actually distort the model. The point has been dealt with in detail in TN 20.

It is possible, then, to conceive of a variety of models for the error term in Equation 1. For one type of error, see the Cicchetti and Smith in an appendix to this volume. Error might be simulated as an interaction effect by taking a certain value when both of two interacting traits are present, and zero otherwise. The error term due to supply and learning factors might be appropriately modelled by the conventional regression error term, an independent, normally distributed variate of zero mean. The latter approach was taken in the simulations which follow.

While the range of empirical p values is clearly restricted to between 0 and 1, there is no explicit requirement that the coefficients of the model be selected so that all predicted values of p lie in that range. In simulation experiments, it is quite likely that p will be driven above 1 or below 0 by the addition of large error terms.

In calibrating the model in other CORD studies this problem has been dealt with by the use of a modified regression procedure which restricts predicted p values to the prescribed range. In these simulations the range has been restricted by truncating any error which would otherwise have driven a p value above 1 or below 0. But in the long term, the problem would be better dealt with by a respecification of the model. Suppose that Equation 1 were to become:

$$p = (1/\Pi) \arctan (a_0 + \Sigma a_i x(i) + \acute{\varepsilon}) + 1/2$$

Then the limits p = 0 and p = 1 would become asymptotes such that even the most favourable or unfavourable combinations of trait variables never quite result in inevitable behaviour, and no restrictions are placed on the values of the $a_i$. Calibration of the specified model is more difficult, because in the appropriate linearized form all values of the dependent variables y become ± infinity. But it would be quite possible to calibrate the model in the non-linear form given above by an appropriate iterative procedure.

SIMULATIONS (1)

The simulations were made with a set of nine independent, binary socioeconomic trait variables. Each one was simulated by generating a uniformly distributed random number in the interval 0 to 1, and then rounding to an integer, so that in each case the probability of the trait being present was 1/2.

p values were calculated from Equation 1. The $\varepsilon$ values were normally distributed with mean 0 and with standard deviation $\sigma$, determined for each simulation, so that the amount of structural error could be varied freely. The method of Box and Muller (1958) was used to generate normally distributed deviates from pairs of random numbers in the interval 0 to 1.

y values were generated from the p's according to the value of a further uniformly distributed random number in the interval 0 to 1. If this number was greater than p, y was set to 0, and otherwise to 1.

The first set of simulations demonstrates the dependence of $R^2$ on statistical error when structural error is absent. The results are shown in Table 1. In each case the fitted vector of coefficients was compared to the original set used to calculate p values. The deviations were expressed in standard errors, and the table shows the mean absolute deviation and the standard deviation of the coefficient deviation, for each run. According to standard regression assumptions, the expected values are 0.798 and 1.0 respectively.

The distribution of p, and the moments kl and k2 are readily calculated, since p is the sum of the constant $a_0$, and nine equally weighted binomial variates $x_1 - x_9$. The probability that exactly r of these variates have the value 1 and the rest 0 is given by the binomial distribution:
$(n!/r!(n-r)!)p^r(1-p)^{n-r}$
WHERE n is the number of binomial variates, and

p is the probability that any one has the value 1.

The mean of the distribution is simply np, and the variance np(1-p) so that:
k1 = 9 x 1/2 x 0.1 + 0.1 = 0.55
$k2 - kl^2$ = 9 x 1/2 x 1/2 x 0.1 x 0.1 x 0.1 = 0.0222
$R^2$ = 0.0222/(O.55 - $0.55^2$) = 0.0898

This is the asymptotic value of $R^2$, to which the values in the table tend as sample size increases.

THE STRUCTURAL COMPONENT

The structural error component can now be reintroduced. Each individual value of p is distorted by a random quantity *e*, which is assumed to have a Normal distribution with a mean 0 and standard deviation $\sigma$.

When structural errors are included, the sum of squares about the regression line becomes:

$$\sum(y-p)^2 = \int(1-p)^2 N(p+e) f(p)dp + \int p^2 N(1-p-e)f(p)dp$$
$$= \int N f(p) (p+e-p^2-2pe)dp$$

(integrating from 0 to 1)

Since only linear terms in *e* appear in the result, it follows that for moderate values of the $\Sigma$ the existence of structural error will have no effect on $R^2$. The proportion of variance explained by the model will be a function of statistical error alone, and will not reflect any structural deficiencies in the analysis.

In the second set of simulations the statistical influence was held constant, and the amount of structural error varied by changing the standard deviation of *e*. The prediction that $R^2$ would remain constant for moderate values of $\Sigma$ is borne out by Table 2. The two measures based on a comparison of the fitted coefficients with the original simulation values show a quite systematic deterioration in the model's ability to recover coefficients as structural error increases. However, there is no such systematic effect on $R^2$, which remains significant at the 95% level on the standard test and is as close to the asymptotic value of 0.0898 when $\Sigma$ = 1 as it is when $\Sigma$ = O. The inevitable conclusion is that low $R^2$ values are a result of the nature of the dependent

variable and do not necessarily indicate any structural weakness in the model.

It is clear from Table 1 that the statistical error component behaves in a similar manner to the error in a standard regression model. The two measures of coefficient recovery remain adequately close to their expected values. However structural error behaves rather differently. Increasing values of $\Sigma$ cause a deterioration in the coefficient measures, but without a corresponding change in $R^2$. The result is that the coefficient measures show a trend to values that are well outside the expected range under standard regression assumptions. In short, the standard error of a regression coefficient can be calculated by the standard methods when statistical error is present, but increasing structural error results in increasing underestimation.

FURTHER ANALYSIS

The analysis thus far has raised two major points. First, the success of the model as expressed in Equation 1 cannot possibly be assessed by means of $R^2$. Second, measures of the standard error in regression coefficients are biased when structural and statistical error are both present. So consider here that the preceding analysis is pursued in more formal terms leading to the discussion of steps that can be taken to rectify the two difficulties just cited.

The fundamental problem can be expressed quite succinctly in terms of the ability of the y values to give accurate estimates of the properties of the p values. Taking the mean first, we have:

$\hat{y} = \int P(y=l|p)\, P(p)dp = \int pP(p)dp = \hat{p}$

Thus the mean of y is an unbiased estimate of the mean of p. Now consider the second moment of y.

$\int P(y=1|p)\, P(p)dp = \hat{p}$

Thus the second moment of y is not an estimate of the second moment of p. It follows that measures based on the sum of squares of p, such as $R^2$, the standard error of regression coefficients, or statistical tests of the significance of $R^2$, will be distorted when y is used as the dependent variable.

TABLE 1: STRUCTURAL ERROR ABSENT ( $\Sigma = 0$)

| Sample Size | $R^2$ | Mean Absolute Error in a | Standard Deviation of Error in a |
|---|---|---|---|
| 20 | .4832 | .943 | 1.048 |
| 50 | .2711 | .459 | .610 |
| 100 | .1085 | .423 | .611 |
| 200 | .0863 | 1.01 | 1.216 |
| 500 | .1236 | 1.07 | 1.277 |

TABLE 2: VARYING STRUCTURAL ERROR

| Sample Size (N) | $\Sigma$ | $R^2$ | Mean Absolute Error in a | Standard Deviation of Error in a |
|---|---|---|---|---|
| 200 | 0 | .0986 | .490 | .600 |
| 200 | .1 | .1586 | .784 | .971 |
| 200 | .5 | .0528 | .927 | 1.147 |
| 200 | 1.0 | .0831 | 1.730 | 2.059 |

Now the crossproduct of y with an independent variable:

$\iint x_i\, P(x_i)dx_i\, P(y=1|p)\, P(p)dp$

$= \int\int x_i\, P(x_i)dx_i pP(p)dp$

which is the crossproduct of p and $x_i$. So the regression coefficient estimates, which are based on the covariances and the variances of the $x_i$, will be unbiased.

Stated in these terms, the problems stem from the impossibility of estimating the variance in values of p. When structural error is assumed absent, the variance is simply that of the fitted p values, so no problem arises, but actual p values may be distorted by the unknown component *e*.

A NEW MEASURE OF STRUCTURAL ERROR

The only possible means of assessing the amount of structural error in the model is by a comparison of the actual and fitted p values. While the observed y values reflect the actual, but unknown p values, they do so only in a statistical sense. In a complex problem, each individual probably possesses a unique combination of independent variable values, or socioeconomic traits, and thus a unique p value. Thus each y value represents a single trial of a unique experiment. (See the review of this chapter for a different comparison method than the one presented here based on the GLS regression methods described in the Cicchetti and Smith appendix to this volume.)

Suppose that the fitted p values were grouped into ranges, say 0 - 0.1, 0.1 - 0.2, etc. Then in the absence of structural error, the proportion of those individuals whose fitted p values lay in each range should be roughly equal to the central p value of each range. Thus approximately 5% of those individuals with a fitted p value in the first range should be observed to participate in the activity. The precise number will be governed by the binomial distribution so that the probability that precisely r individuals will participate is:

$P(r) = ((n-r)!/r!)P(c)^r (1-P(c))^{(n-r)}$ WHERE the central p value is P(c), and n individuals have p values in the range.

Unfortunately, if we assume that structural error distorts p values in a normal distribution with a mean of 0, then the proportion of individuals will remain roughly the same, irrespective of the amount of structural error present, since upward distortions in p are as likely as downward distortions. However, it is clear that towards the limits of 0 and 1, the normal distribution is not a reasonable model of structural error, since near 0 upward distortions must be more likely, and conversely near 1. This suggests, then, that the deviations in observed proportions from the central p values in each range, particularly near 0 and 1 may be a reasonable measure of structural error.

These ideas are now clarified with an example. Figure 1 shows the distribution of fitted p values for the simulation run s = 0, N = 200, that is, a sample of 200 generated with no structural error. Figure 2 shows the distribution of numbers of participants by fitted p value, so that each bar represents the number of individuals who participated, and who had fitted p values in that range. Table 3 shows the corresponding proportions.

The reliability of each observed proportion depends on the p value count in that interval. Clearly values toward the middle of the table are based on larger samples and should thus be given greater weight. In view of this, a better strategy might be to organize the distribution into intervals of equal numbers of observations. Table 4 shows the same distribution in the form of the ten deciles, so that each interval contains precisely 20 observations of fitted p values.

The proposed structural error index S(2) is based on the fit of observed proportions to central p values, when the data is arranged by deciles, weighted by the expected reliability of each proportion in a modified $R^2$ statistic:

$$S^2 = 1 - \frac{\Sigma_i (s_i - x_i)2/(x_i - x_i^2)}{\Sigma_i (s_i - 0.5)^2 /(x_i - x_i^2)}$$

WHERE $s_i$ = observed proportion in the $i^{th}$ decile and x(i) is the central p value, and i=1 to 10.

TABLE 3: FITTED p DISTRIBUTION

| Central p | p Value Count | Participant Count | Proportion | Standard Error |
|---|---|---|---|---|
| .05 | 1 | 1 | 1.0 | .22 |
| .15 | 8 | 4 | .5* | .13 |
| .25 | 15 | 0 | 0 | .11 |
| .35 | 31 | 13 | .42 | .09 |
| .45 | 37 | 16 | .43 | .08 |
| .55 | 42 | 16 | .38 | .08 |
| .65 | 34 | 28 | .82 | .08 |
| .75 | 23 | 18 | .78 | .09 |
| .85 | 8 | 6 | .75 | .13 |
| .95 | 1 | 1 | 1.0 | .22 |

* Probability of observing a value at least this different from the central p is less than 5%.

TABLE 4

FITTED p DISTRIBUTION BY DECILES

| Central p | p Value Count | Participant Count | Proportion | Std Error |
|---|---|---|---|---|
| .151 | 20 | 8 | .40* | .08 |
| .333 | 20 | 4 | .20 | .11 |
| .386 | 20 | 5 | .25 | .11 |
| .431 | 20 | 12 | .60 | .11 |
| .470 | 20 | 6 | .30 | .11 |
| .506 | 20 | 11 | .55 | .11 |
| .553 | 20 | 10 | .50 | .11 |
| .610 | 20 | 10 | .50 | .11 |
| .668 | 20 | 16 | .80 | .11 |
| .802 | 20 | 17 | .85 | .09 |

* Probability of observing a value at least this different from the central p is less than 5%.

Table 5 shows the computed values for the index for the conditions used in Table 2. Each experiment was repeated 25 times, and the table shows the observed mean and standard error of $S^2$.

The index shows the expected trend. The upper limit of the scale, corresponding to $\Sigma = 0$, depends on the sample size and will approach 1 as the sample increases and observed proportions become more accurate estimates of central p values.

SUMMARY

This paper has been concerned with models which relate an individual's socioeconomic characteristics or traits to the probability of his participating in specific recreation activities. Such models are calibrated by using as the dependent variable a binary representation of the individual's actual behaviour, which is only statistically related to the model probability.

Under these conditions, the $R^2$ statistic is determined by the statistical relationship between probability and binary event, and in no way reflects the model's actual goodness of fit, or its structural error. Estimates of regression coefficients are unbiased, but since the binary event variable cannot give an estimate of the variance of the probability variable, $R^2$ and related statistics are of questionable value.

TABLE 5 EFFECT OF VARYING Σ ON $S^2$ (N = 200)

| Σ | Mean $S^2$ | Standard Error |
|---|---|---|
| 0 | .782 | .120 |
| .1 | .766 | .147 |
| .5 | .599 | .222 |
| 1.0 | .616 | .187 |
| 2.O | .271 | .310 |

If each individual must be assumed to have a unique socioeconomic mix, then each event is a trial under a different probability, and measurement of errors in those probabilities is extremely difficult. The method suggested is a measure of the goodness of fit between probabilities and proportions of events when observations are grouped into ranges of probabilities. It is argued that structural error will produce distortions in this fit towards the ends of the probability spectrum at 0 and B.

Simulations were used to demonstrate the proposed index. They show the expected behaviour under increasing amounts of structural error. Standard errors are quite large, and the expected values of the index depend on sample size, on the form of structural error, and also on the parameters of the model, so that the index should be used only as a relative scale and not as an absolute measure. It would be wise to simulate the expected values of the index under the specific conditions of each application.

Should the model be applied to cases where numbers of individuals have the same socioeconomic characteristics, many of the problems are much less severe. Calibration can be carried out using observed proportions as the dependent variable rather than binary events. The reliability of each proportion depends on the number of individuals involved, so that the data should be weighted accordingly. The problem has been discussed in the context of recreation flow models in TN No. 19, and many of the conclusions of that paper can be applied to the participation modelling problem.